Assessment of Model Drifts in Seasonal Forecasting: Sensitivity to Ensemble Size and Implications for Bias Correction

Rodrigo Manzanas

Meteorology Group
Dpto. de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria
Santander, 39005, Spain

Key Points:

- Model continues to drift well beyond the first month after initialization, leading to significant space-and-time varying drifts
- The ensemble size needed to robustly characterize model drifts depends on the underlying predictive skill
- Moving windows can help to remove the unwanted effects coming out from the model drift, which can lead to important intra-seasonal biases

Corresponding author: Rodrigo Manzanas, rodrigo.manzanas@unican.es

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2019MS001751

Abstract

Despite its systematic presence in state-of-the-art seasonal forecasts, the model drift (leadtime-dependent bias) has been seldom studied to-date. To fill this gap, this work analyzes its spatio-temporal distribution, and its sensitivity to the ensemble size in temperature and precipitation forecasts. Our results indicate that model continues to drift well beyond the first month after initialization, leading to significant, highly space- and time-varying drifts over vast regions of the world. Nevertheless, small ensembles (less than 10 members) are enough to robustly estimate the mean model drift and its year-to-year fluctuations in skillful regions. Differently, in regions of low model skill, larger ensembles are required to appropriately characterize this inter-annual variability, which is often larger than the drift itself. This points out a necessity to develop new strategies which allow for efficiently dealing with model drift, especially when bias correcting seasonal forecasts —most of the techniques used to this aim rely on the assumption of stationary model errors.— We demonstrate here that the use of moving windows can help to remove not only the mean forecast bias, but also the unwanted effects coming out from the drift, which can lead to important intra-seasonal biases if it is not properly taken into account. The results from this work can help to identify the nature and causes of some of the systematic errors in current coupled models, and can have large implications for a wide community of users who need long, continuous unbiased seasonal forecasts to run their impact models.

Plain Summary Language

This work analyzes the satio-temporal distribution of the model drift (leadtime-dependent bias), as well as its sensitivity to the ensemble size in the context of seasonal forecastig. The results obtained indicate that model continues to drift well beyond the first month after initialization, leading to significant, highly space- and time-varying drifts over vast regions of the world. Nevertheless, small ensembles (less than 10 members) are enough to robustly estimate the mean model drift and its year-to-year fluctuations in skillful regions. In addition to this, this paper demonstrates that the use of moving windows can help to remove not only the mean forecast bias, but also the unwanted effects coming out from the drift, which can lead to important intra-seasonal biases if it is not properly taken into account. These results can have large implications for a wide community of users who need long, continuous unbiased seasonal forecasts to run their impact models.

1 Introduction

Seasonal forecasts have enormous impact on different socioeconomic sectors such as agriculture, tourism, energy and health [see, e.g., *Hill and Mjelde*, 2002; *Doblas-Reyes et al.*, 2013, and references therein]. Nowadays, these forecasts are routinely produced (and delivered) by a number of WMO-designated Global Producing Centres for Long Range Forecasts (GPCs: http://www.wmo.int/pages/prog/wcp/wcasp/gpc/gpc.php), based on different state-of-the-art global ocean-atmosphere coupled models. However, due to the important simplifications that need to be done when building these models —which can lead to deficient representations of circulation, energy exchanges, etc.,— seasonal forecasts are known to present important errors, either at regional or local scales [see, e.g., *Manzanas et al.*, 2019, 2018, respectively]. In particular, in addition to the systematic mean error or standard bias (mean deviation from observations for a particular target period and location), a second order bias which depends on the leadtime —the time that passes from the moment in which the model is initialized to the start of the target period to be predicted— arises in seasonal forecasting. The latter, known as drift, is a consequence of having initial conditions inconsistent with the model dynamics [*Alves et al.*, 2004; *Fernández et al.*, 2009], and can be defined as the tendency of the model to evolve from the initial (observed) state to its own attractor [see, e.g., *Delworth et al.*, 2006; *Collins et al.*, 2006; *Magnusson et al.*, 2013; *Doblas-Reyes et al.*, 2013]. This tendency, which should

not be confused with a seasonal climate signal, can lead to important leadtime-varying errors which can considerably affect the quality of the forecasts [*Smith et al.*, 2013; *Vannitsem et al.*, 2018]. In this aspect, many previous works have documented substantial errors in key fields such as precipitation when the coupled models are initialized with observed SSTs [see, e.g., *Troccoli et al.*, 2008]. Moreover, the origin and causes responsible for model drift are not obvious, and it is highly dependent on the variable and the geographical area analyzed [*Bedia et al.*, 2018].

Despite all this, only a few studies have paid attention to the issue of model drift in the context of seasonal forecasting [see, e.g., *Stockdale*, 1997], being most of the previous works focused on decadal predictions, and more concretely, on finding ways of reducing the long-term drift by performing some kind of correction on the initial conditions [see, e.g., *Zhang*, 2011; *Kharin et al.*, 2012; *Zhang et al.*, 2012; *Fučkar et al.*, 2014; *Sánchez-Gomez et al.*, 2016]. At seasonal time-scales, *Shonk et al.* [2018] found a tendency in the simulated ITCZ to move to the north (as compared to observations) —they refer to this effect as a spatial drift.— More recently, *Hermanson et al.* [2018] analyzed model biases and drifts for different time-scales, including both decadal but also seasonal forecasts. Even though this work provides essential knowledge to better understand the model drift, they only focused on a few regions (with extratropical latitudes being misrepresented) and only assessed average model drifts, without worrying about their temporal variability or the importance of the ensemble size in characterizing those drifts. Moreover, they did not propose any strategy for correction.

Therefore, to overcome the existing necessity of providing a complete diagnosis of model drifts globally [*Hermanson et al.*, 2018], this work analyzes their spatio-temporal distribution for seasonal forecasts of temperature and precipitation worldwide. This can help to identify specific model deficits and offers the possibility of targeted improvement of certain processes formulation, resolution and parametrization [*Ehret et al.*, 2012]. Nevertheless, it is important to note that identifying/understanding the mechanisms of the physical processes involved in the model drift is out of the scope of the present study — for this particular regard, the reader is referred to a few relevant previous papers on the topic [see, e.g., *Magnusson et al.*, 2012, 2013; *Carrassi et al.*, 2014].— Furthermore, there exists an open discussion on the ensemble size that is required for a proper statistical correction of model errors in seasonal forecasting. Using as many members as possible is usually the preferred option. However, *Manzanas et al.* [2019] have recently shown that small ensembles are enough to robustly correct the mean model biases. We test here if this also holds for the case of model drift.

Finally, we explore the suitability of considering moving windows [*Bedia et al.*, 2018] for the application of a standard quantile-mapping technique as a way to minimize the unwanted effects coming out from model drifts —note that most of the state-of-the-art techniques that are used for bias correction of raw seasonal forecasts [see *Manzanas et al.*, 2019, for a review] rely on the assumption of stationary model errors, and their application in non-stationary circumstances remains unclear [see, e.g., *Anderson*, 2011].— To do this we focus on the Philippines, where most important sectors could greatly benefit from the use of suitable, unbiased seasonal forecasts.

The paper is organized as follows: The data and the methodology used are described in Section 2. Results are presented through Section 3. The most important conclusions are given in Section 4.

2 Seasonal Forecasts and Definition of Drift Used

Daily temperature and precipitation from the European Center for Medium Weather Forecasts (ECMWF) System 4 [*Molteni et al.*, 2011] were considered over the entire globe at their original spatial resolution (0.75°). System 4 is based on the atmospheric model

IFS (cycle 36r4) and the oceanic model NEMO, and provides the longest-to-date seasonal hindcast, covering the period 1981-2010, which is fully used here. Whereas IFS is initialized with ERA-Interim [*Dee et al.*, 2011] data, the ocean data assimilation system NEMOVAR is used to initialize NEMO. By producing a set of perturbed initial conditions, an ensemble of 15 members is generated —in particular, five members originate from perturbations of ocean wind surface initial conditions, whereas the other 10 members originate from sea surface temperature perturbations and stochastic physics— the first day of each month and run for 7 months. Note therefore that there are seven possible leadtimes for predicting each calendar target month. For instance, mean conditions for August can be computed considering the forecasts initialized the first of that same August (i.e., leadtime 0), the previous July (leadtime 1) and so on until February (leadtime 6) —leadtime is expressed in months in the forthcoming.—

For a particular gridbox and target month M, the mean drift $\overline{d}$ is defined as the difference between climatologies $\overline{p}$ —as given by the ensemble mean over 1981-2010— of M at different leadtimes (see Figure 1). For instance, for leadtimes LT_i and LT_j (assume $j > i$):

$$\overline{d}_{M_{LT_j},LT_i} = \overline{p}_{M_{LT_j}} - \overline{p}_{M_{LT_i}} \qquad i, j = 0, ..., 6 \tag{1}$$

Note the convenience of this way of defining the drift, since it does not involve the use of any verification data, eliminating thus the issue of observational uncertainty [*Kotlarski et al.*, 2017; *Herrera et al.*, 2018] —different results might come out if the drift was defined relative to observations.—

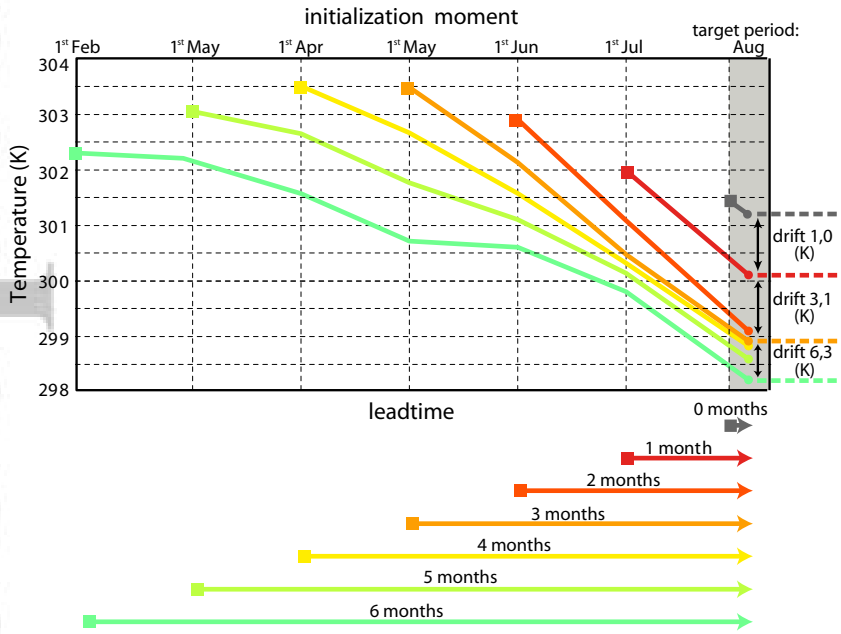


Figure 1. Schematic illustration of the model drift —as defined in this work— for the particular case of mean predictions for August, which can be obtained from the initialization of the 1st of August (i.e. at leadtime 0), July (1-month leadtime), June (2-months leadtime), etc.

3 Results

3.1 Statistical Assessment of Model Drifts

Figure 2 shows the mean value of the drift (Equation 1) for temperature (top) and precipitation (bottom). For brevity, results are only shown for four illustrative target months: February, May, August and November. Only drifts significantly ($\alpha = 0.05$) different from zero are displayed —white gridboxes identify areas where not significant values were found.— To compute the significance of the drift, a bootstrap approach was followed [*Mason and Graham*, 2002]. In particular, 1000 different ensembles of 15 members each were built by random selection amongst the 15 members available (allowing for repeated members). The mean drift and its confidence intervals were then computed upon the 1000 bootstrapped values, which were derived from the 1000 different ensembles. Note also that drifts are only shown for a selection of incrementally increasing leadtimes. In particular, the left/middle/right column corresponds to the drift between leadtimes 1 and 0/3 and 1/6 and 3. We do this in order to properly assess how the drift varies along the entire model run.

Although providing a detailed description of the patterns found is not the aim here, there are some important conclusions which must be mentioned. First, significant drifts are found worldwide (especially for temperature). Second, model drifts considerably vary both in space and time. Third, despite the drifts shown in the first column might be expected to be larger than those in columns 2 and 3 due to the rapid adjustment processes that occur during the first days of the model run when the atmosphere and the ocean are initialized from different systems [see, e.g., *Balmaseda*, 2012], this figure shows that the model continues to drift significantly well beyond the first month after the initialization moment. This is particularly the case for temperature, for which the model drifts are even larger during the last part of the run. For this variable, drifts are mostly negative and are present over vast parts of the globe (with absolute values exceeding 1.5 K in some regions). Moreover, they are stronger over land (especially in large portions of the northern hemisphere such as North America and Siberia) than over the oceans. Differently, for precipitation, drifts are mainly located in tropical latitudes and are more pronounced over the oceans (with absolute values above 50 $mm/month$ in many cases). Note that, in some regions, these drifts can be close to or larger than the underlying climatology, which may seriously reduce potential model skill [*Smith et al.*, 2013].

For a better characterization of the drifts shown in Figure 2, we also analyzed their inter-annual variability. In particular, Figure 3 shows the standard deviation of the year-to-year drifts —which is simply referred to as $\sigma(d)$ hereafter— divided by their mean absolute value (see Figure 2), for temperature (top) and precipitation (bottom). Therefore, green (brown) colors identify those gridboxes where $\sigma(d)$ is smaller (larger) than the mean drift itself. In other words, brown colors correspond to regions where the model drift is far from being stationary.

According to this figure, temperature drifts might be only safely removed in the tropics (green areas) under the assumption of stationary model errors. The situation is even more problematic for precipitation, for which brown colors are predominant in most regions (with some exception as the Gulf of Guinea), and therefore, time-invariant corrections would not be optimal to correct the year-to-year drifts. These results suggest that there is a need to develop new strategies which allow for efficiently dealing with model drifts. In this regard, we test in Section 3.3 if the use of moving windows can help to remove not only the mean bias from raw seasonal forecasts, but also the unwanted effects that may appear due to the model drift.

3.2 Sensitivity to Ensemble Size

Given the high computational costs involved in the generation of large ensembles for seasonal forecasting, it is important to have some estimation about the number of members (n in the following) that are needed to robustly characterize particular model properties, for instance the drift. Therefore, Figure 4 shows how the mean drift obtained for temperature and precipitation (left and right column respectively) varies with n for the case of February (similar results are obtained for the rest of months). The analysis is performed for two illustrative regions —the spatially averaged time-series are considered,— Europe and El Niño 3 (shadowed areas in the top and bottom row, respectively). Note that whereas low-to-moderate seasonal predictive skill is in general acknowledged for the former, high model skill has been documented in the latter [see, e.g., *Manzanas et al.*, 2014]. Ensembles of increasing size ($n = 1, ..., 15$) were built based on bootstrapping [*Mason and Graham*, 2002]. In particular, for each n, 1000 different ensembles was constructed by randomly selecting n members out of the 15 available ones (repeated members are allowed). From these 1000 ensembles, the mean drift value and its standard deviation (errorbars) were obtained.

Interestingly, this figure reveals that the mean drift is almost independent of n, and therefore, small ensembles would be sufficient for a robust estimation of model drifts. For instance, the drift patterns obtained for $n = 5$ are very similar to those presented in Figure 2 (not shown).

Additionally, we also assessed how the standard deviation of the year-to-year drift, $\sigma(d)$, depends on n. Taking into account that $\sigma^2(ax - by) = a^2\sigma^2(x) + b^2\sigma^2(y) - 2ab \cdot cov(x, y)$ for any two samples from random variables x and y and two scalars a and b — where σ^2 is the variance and cov the covariance operator,— note from Equation 1 that:

$$\sigma^2(d_{M_{LT_j,LT_i}}) = \sigma^2(p_{M_{LT_j}} - p_{M_{LT_i}}) = \\ = \sigma^2(p_{M_{LT_j}}) + \sigma^2(p_{M_{LT_i}}) - 2 \cdot cov(p_{M_{LT_j}}, p_{M_{LT_i}}) = \tag{2}$$

Taking also into account that $cov(x, y) = \sigma(x) \cdot \sigma(y) \cdot \rho(x, y)$ —where ρ is the correlation coefficient,— this can be expressed as

$$\sigma^2(d_{M_{LT_j,LT_i}}) = \\ = \sigma^2(p_{M_{LT_j}}) + \sigma^2(p_{M_{LT_i}}) - 2 \cdot \sigma(p_{M_{LT_j}}) \cdot \sigma(p_{M_{LT_i}}) \cdot \rho(p_{M_{LT_j}}, p_{M_{LT_i}}) = \tag{3}$$

where $\sigma^2(p_{M_{LT_j}})$ and $\sigma^2(p_{M_{LT_i}})$ represents the variance of the year-to-year ensemble mean predictions, p, at leadtime j and i, respectively. However, according to a Bartlett test [*Snedecor and Cochran*, 1989] —which checks the null hypothesis of equal variances across different samples,— we found that $\sigma^2(p_{M_{LT_1}}) \simeq \sigma^2(p_{M_{LT_3}}) \simeq \sigma^2(p_{M_{LT_6}})$ in about the 99% of global gridboxes, either for temperature or for precipitation, and for all target months (not shown). The only leadtime for which the null hypothesis of the Bartlett test can be rejected ($\alpha = 0.05$) in a considerable number of gridboxes is 0, which is due to the aforementioned rapid adjustment processes that occur during the first days of the model run in seasonal forecasting. Therefore,

$$\sigma(p_{M_{LT_i}}) \simeq \sigma(p_{M_{LT_j}}) \equiv \sigma(p_M) \\ \forall i, j \neq 0 \tag{4}$$

and Equation 3 can be rewritten as:

$$\sigma^2(d_{M_{LT_j,LT_i}}) = 2\sigma^2(p_M)\left[1-\rho(p_{M_{LTi}}, p_{M_{LTj}})\right] \quad \forall i,j \neq 0 \tag{5}$$

And, finally:

$$\sigma(d_{M_{LT_j,LT_i}}) = \underbrace{\sqrt{2}\sigma(p_M)}_{term\ (1)} \underbrace{\left[1-\rho(p_{M_{LT_i}}, p_{M_{LT_j}})\right]^{\frac{1}{2}}}_{term\ (2)} \tag{6}$$

which indicates that the inter-annual variability of the drift comes determined by a first term which is basically $\sigma(p_M)$ (the inter-annual variability of the ensemble mean at any leadtime except 0), and a second term which is related to the persistence of the model. With respect to the latter, note that high values of ρ would reflect that the model provides consistent (similar) predictions independently of the leadtime considered. It is reasonable to think that such situations will occur in cases for which a persistent predictability signal exists, and thus, skillful predictions could expected.

Figure 5 puts some light on the contribution of each of these two terms for the case of temperature over Europe and El Niño 3 for February (similar conclusions are obtained for the rest of months, as well as for precipitation). In particular, the left column displays the de-trended (Mann-Kendall test with $\alpha = 0.05$) year-to-year predictions for the ensemble mean and a single —randomly selected— member (solid and dashed lines, respectively). Black (blue) represents one-(three-) month lead predictions —similar results are found for the rest of leadtimes.— The right column shows the corresponding year-to-year drifts.

In Europe (El Niño 3), $\sigma(p_M) << \sigma(p_M^{1memb})$ ($\sigma(p_M) \simeq \sigma(p_M^{1memb})$) —see the numbers in the upper corners inside the panels,— which reflect a low (high) inter-member consistency leading to low (high) correlations between predictions at leadtime 1 and 3, either for the ensemble mean or for the aleatory member —see the numbers in the lower right corner.— Therefore, according to Equation 6, and taking into account that $\sigma(\overline{x}) = \frac{1}{\sqrt{N}}\sigma(x)$ —where N is the sample size of the random variable x,— the inter-annual variability of the drift should decay with n as $1/\sqrt{n}$ —although modulated by (1) and (2)— in regions of low model skill such as Europe,

$$\sigma(d_{M_{LT_j,LT_i}}) = \sqrt{\frac{2}{n}}\sigma(p_M^{1memb})\left[1-\rho(p_{M_{LT_i}}, p_{M_{LT_j}})\right]^{\frac{1}{2}} \tag{7}$$

whereas it should be solely determined by (1) and (2) in skillful regions such as El Niño 3.

$$\sigma(d_{M_{LT_j,LT_i}}) = \sqrt{2}\sigma(p_M^{1memb})\left[1-\rho(p_{M_{LT_i}}, p_{M_{LT_j}})\right]^{\frac{1}{2}} \tag{8}$$

To check this premise, Figure 6 shows the standard deviation of the year-to-year drifts as a function of n for the same example of Figure 5. Again, for each n (with $n = 1, ..., 15$), 1000 different ensembles were constructed by randomly selecting n members out of the 15 available ones (allowing for repeated members). The mean value (solid line) and the standard deviation (errorbars) of the drift were obtained from these 1000 bootstrapped ensembles. Additionally, dashed (dotted) lines draw the theoretical Equation 7 (8) for regions of low (high) model skill.

The experimental results are very close to the theoretical expected ones, especially in El Niño 3. Yet, in this region, the inter-annual variability of the drift decays with n —especially for the first few members,— which implies that either $\sigma(p_M^{1memb})$ or $\rho(p_{M_{LT_i}}, p_{M_{LT_j}})$ should vary with n. Indeed, Figure 7 evidences that $\rho(p_{M_{LT_i}}, p_{M_{LT_j}})$ increases with n (markedly for the few first members) in El Niño 3, which would explain the decaying graph obtained in Figure 6 for this region. Moreover, Figure 7 also confirms that, in Europe, the inter-annual variability of the drift exclusively depends on n, since neither $\sigma(p_M^{1memb})$ nor $\rho(p_{M_{LT_i}}, p_{M_{LT_j}})$ fluctuate with n.

In summary, our results indicate that small ensembles ($n < 10$) are enough to robustly estimate the mean value —computed over a sufficiently long period— of model drifts, independently of the variable and/or the region being considered. However, whereas such small ensembles also allow to capture the representative year-to-year fluctuations of these drifts in skillful regions, larger ensembles are required to this aim in regions of low model skill. Analysis such as the one undertaken here may help to determine the optimum number of members needed in the different regions of the world.

3.3 Implications for Bias Correction

In order to better understand the role that the model drifts shown so far may play when standard state-of-the-art techniques are used to bias correct raw model seasonal forecasts, we focus on the Philippines, a moderately skillful region for which this type of predictions is key for various sectors —e.g. rice production [*Koide et al.*, 2012].— In particular, we used the 42 gauge stations made available by the Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA: http://www.pagasa.dost.gov.ph), which cover the four climatic types present in the country [*Coronas*, 1920; *Flores and Balagot*, 1969; *Kintanar*, 1984] and have been already used in previous studies [see, e.g., *Manzanas et al.*, 2015, 2018]. Instead of the usual 3-month long seasons (DJF, MAM, JJA, SON) we focus here on extended 6-month long ones (DJFMAM, MAMJJA, JJASON, SONDJF), which allow for better illustrating the unwanted effects introduced by model drifts. One-month lead predictions of temperature and precipitation from the ECMWF System4 were first bi-linearly interpolated to the 42 PAGASA stations. Then, an empirical quantile–mapping method participating in the VALUE downscaling intercomparison initiative [*Maraun et al.*, 2017] —referred to as EQM hereafter— was applied to correct them. This method, which has been recently applied in the context of seasonal forecasting [see, e.g., *Manzanas et al.*, 2018; *Manzanas and Gutiérrez*, 2018], consists of calibrating the predicted empirical probability density function by adjusting a number of quantiles based on the empirical observed one [see, e.g., *Déqué*, 2007]. In particular, for each gridbox, we adjusted percentiles 1 to 99 and linearly interpolated every two consecutive percentiles inside this range. Outside, a constant extrapolation (using the correction obtained for the 1st or 99th percentile) was applied. All members were independently corrected based on their joint distribution. For the case of precipitation, the frequency adaptation proposed by *Themeßl et al.* [2012] was applied to account for possible cases for which the predicted frequency of dry days was larger than the observed one. To avoid over-fitting, a leave-one-out cross validation scheme [*Lachenbruch and Mickey*, 1968] in which each year was separately considered for test whilst keeping the rest for training was applied. Beyond the standard implementation of the method, in which the total test set in each iteration is corrected at once (based on the total train set available), we also assessed here the suitability of using moving windows, which allow for independently correcting determined consecutive periods (e.g. days, weeks), based on a collection of data centered around the target period being corrected. For instance, if the method is applied on a daily basis, and we want to correct the forecast for 16-Jan-1981, we may use all January days from the period 1982-2010 for the mapping. This configuration would correspond to a 31-day window, and it is the one used here. This choice for the width of the window is based on *Bedia et al.* [2018], who found, in the context of seasonal forecasting, that a 31-day window as-

sures for smooth daily transitions whilst being narrow enough to encompass periods for which the possible trends introduced by model drift can be safely neglected.

Figure 8 shows, for the illustrative case of temperature, the mean bias obtained for the different extended seasons (in columns) at the 42 PAGASA stations, as given by the raw model forecasts (top row) and a standard EQM in which moving windows are not considered (bottom row).

As expected by construction, the biases are basically negligible after applying the EQM method. However, if these results are individually analyzed for each month within the season, important sign-varying intra-seasonal (i.e. monthly) biases appear as a consequence of the evolving model drifts (see Section 3.1). An example is provided in the top row of Figure 9 for MAMJJA.

In this case, the use of moving windows in the application of the EQM method virtually eliminates the unwanted effects of model drift, leading to negligible biases for all months within the season (bottom row). Moreover, as shown in Figure 10, this also holds for the rest of seasons and for precipitation Nevertheless, despite their suitability to correct the intra-seasonal biases, it must be also noticed that moving windows do not allow for improving the interannual skill of the corrected predictions. An example of this can be seen in Figure 11, where the interannual correlation with observations is shown for predictions of monthly temperature given by the EQM method when moving windows are not/are considered (top/botom row). Despite its inability to improve the predictive skill, the results found in this work regarding the application of moving windows might still be key to many sectors which need long, continuous unbiased climate forecasts (e.g. hydrology or crop modelling).

4 Conclusions

Despite its systematic presence in state-of-the-art seasonal forecasts, a rigorous statistical characterization of model drift (leadtime-dependent bias) is still lacking for this type of prediction. To fill this gap, the present work analyzes the spatio-temporal distribution of model drifts for global seasonal forecasts of temperature and precipitation — the most important variables in user's applications— and their sensitivity to the ensemble size.

Our results indicate that model continues to drift well beyond the first month after the initialization moment, leading to significant, highly space- and time-varying drifts over vast regions of the world (especially for temperature). Nevertheless, small ensembles (less than 10 members) are enough to robustly estimate the mean value of these drifts, independently of the variable and/or the region being considered, and also allow to capture their year-to-year fluctuations in regions with good model predictive skill. This important finding suggests that costly approaches for seasonal impacts forecasting (e.g. dynamical downscaling) might benefit from drift removal strategies involving smaller ensemble sizes in skillful regions.

Differently, in regions of low model predictive skill, larger ensembles are required to appropriately characterize the year-to-year variability of model mean drifts, which is detected to be larger than the drift itself in many cases (especially for precipitation). This points out an existing necessity to develop new strategies which allow for efficiently dealing with model drifts, especially when bias correcting seasonal forecasts —note that most of the state-of-the-art techniques used to this aim rely on the assumption of stationary model errors.— In this regard, we demonstrate here that the use of moving windows can help to remove not only the mean bias, but also the unwanted effects coming out from the drift. This is illustrated for the Philippines, where the use of moving windows virtually eliminates the intra-seasonal biases that emerge when entire seasons are corrected at once (as it is usually done). This can have important practical implications for a broad community of users who need long, continuous unbiased seasonal climate forecasts to run

their impact (e.g. hydrology or crop) models. Also, and despite it is out of the scope of this study, the results shown here can help to better understand how the the state-of-the-art coupled models work, and particularly, to identify the nature and causes of some of the most important accompanying errors, which is still a major task in seasonal forecasting. This could be the aim for a future paper.

Finally, it is worth to notice that all the analyses presented in this paper rely on a single forecasting model, the ECMWF System 4. To further test the robustness of the results found, in particular regarding the usefulness of moving windows to provide long, unbiased seasonal forecasts, we plan to extend this study by including a number of newer forecasting systems. This will be the focus for a future paper.

Acknowledgments

This study was supported by the EU projects EUPORIAS (EUropean Provision Of Regional Impact Assessment on a Seasonal-to-decadal timescales) and SPECS (Seasonal-to-decadal climate Prediction for the improvement of the European Climate Services), funded by the European Commission's Seventh Framework Research Programme through grant agreements 308291 and 308378, respectively. The ECMWF System 4 data used here can be retrieved from the User Data Gateway (UDG), a THREDDS-based service from the Santander Climate Data Service which provides access to a wide catalogue of popular climate datasets: http://meteo.unican.es/tds5/catalogs/system4/System4Datasets.html (new users need to register first: http://meteo.unican.es/udg-tap/home). The observational data considered for the Philippines have been privately provided by PAGASA. Please contact directly PAGASA staff (http://bagong.pagasa.dost.gov.ph/aboutus/key-officials) for new requests. Finally, the author wants to acknowledge J. M. Gutiérrez for his constructive comments and continuous support during the elaboration of this work.

References

Alves, O., M. A. Balmaseda, D. Anderson, and T. Stockdale (2004), Sensitivity of dynamical seasonal forecasts to ocean initial conditions, *Quarterly Journal of the Royal Meteorological Society*, *130*(597), 647–667, doi:10.1256/qj.03.25.

Anderson, D. (2011), Current capabilities in sub-seasonal to seasonal prediction. In: Workshop on Sub-seasonal to Seasonal Prediction, *WRCP*.

Balmaseda, M. (2012), Initialization techniques in seasonal forecasts, in *Seminar on seasonal prediction: Science and applications*, ECMWF.

Bedia, J., N. Golding, A. Casanueva, M. Iturbide, C. Buontempo, and J. M. Gutiérrez (2018), Seasonal predictions of Fire Weather Index: Paving the way for their operational applicability in Mediterranean Europe, *Climate Services*, *9*, 101 – 110, doi: 10.1016/j.cliser.2017.04.001.

Carrassi, A., R. Weber, V. Guemas, F. Doblas-Reyes, M. Asif, and D. Volpi (2014), Full-field and anomaly initialization using a low-order climate model: A comparison and proposals for advanced formulations, *Nonlinear Processes in Geophysics*, *21*(2), 521–537, doi:10.5194/npg-21-521-2014.

Collins, W. D., C. M. Bitz, M. L. Blackmon, G. B. Bonan, C. S. Bretherton, J. A. Carton, P. Chang, S. C. Doney, J. J. Hack, T. B. Henderson, J. T. Kiehl, W. G. Large, D. S. McKenna, B. D. Santer, and R. D. Smith (2006), The Community Climate System Model Version 3 (CCSM3), *Journal of Climate*, *19*(11), 2122–2143, doi: 10.1175/JCLI3761.1.

Coronas, J. (1920), *The climate and weather of the Philippines, 1903-1918*, PAGASA.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Holm, L. Isaksen, P. Kallberg, M. Koehler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J. J. Morcrette, B. K.

Park, C. Peubey, P. de Rosnay, C. Tavolato, J. N. Thepaut, and F. Vitart (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597, doi:10.1002/qj.828.

Delworth, T. L., A. J. Broccoli, A. Rosati, R. J. Stouffer, V. Balaji, J. A. Beesley, W. F. Cooke, K. W. Dixon, J. Dunne, K. A. Dunne, J. W. Durachta, K. L. Findell, P. Ginoux, A. Gnanadesikan, C. T. Gordon, S. M. Griffies, R. Gudgel, M. J. Harrison, I. M. Held, R. S. Hemler, L. W. Horowitz, S. A. Klein, T. R. Knutson, P. J. Kushner, A. R. Langenhorst, H.-C. Lee, S.-J. Lin, J. Lu, S. L. Malyshev, P. C. D. Milly, V. Ramaswamy, J. Russell, M. D. Schwarzkopf, E. Shevliakova, J. J. Sirutis, M. J. Spelman, W. F. Stern, M. Winton, A. T. Wittenberg, B. Wyman, F. Zeng, and R. Zhang (2006), GFDL's CM2 Global Coupled Climate Models. Part I: Formulation and Simulation Characteristics, *Journal of Climate*, *19*(5), 643–674, doi:10.1175/JCLI3629.1.

Déqué, M. (2007), Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values, *Global and Planetary Change*, *57*(1-2), 16–26, doi:10.1016/j.gloplacha.2006.11.030.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues (2013), Seasonal climate predictability and forecasting: status and prospects, *Wiley Interdisciplinary Reviews: Climate Change*, *4*(4), 245–268, doi:10.1002/wcc.217.

Ehret, U., E. Zehe, V. Wulfmeyer, K. Warrach-Sagi, and J. Liebert (2012), HESS Opinions "Should we apply bias correction to global and regional climate model data?", *Hydrology and Earth System Sciences*, *16*(9), 3391–3404, doi:10.5194/hess-16-3391-2012.

Fernández, J., C. Primo, A. S. Cofiño, J. M. Gutiérrez, and M. A. Rodriguez (2009), MVL spatiotemporal analysis for model intercomparison in EPS: application to the DEMETER multi-model ensemble, *Climate Dynamics*, *33*(2-3), 233–243, doi:10.1007/s00382-008-0456-9.

Flores, J. F., and V. F. Balagot (1969), *World Survey of Climatology*, *Climates of Northern and Eastern Asia*, vol. 8, chap. Climate of the Philippines, pp. 159–213, Arakawa.

Fučkar, N. S., D. Volpi, V. Guemas, and F. J. Doblas-Reyes (2014), A posteriori adjustment of near-term climate predictions: Accounting for the drift dependence on the initial conditions, *Geophysical Research Letters*, *41*(14), 5200–5207, doi:10.1002/2014GL060815.

Hermanson, L., H.-L. Ren, M. Vellinga, N. Dunstone, P. Hyder, S. Ineson, A. Scaife, D. Smith, V. Thompson, B. Tian, and K. Williams (2018), Different types of drifts in two seasonal forecast systems and their dependence on enso, *Climate Dynamics*, *51*(4), 1411–1426, doi:10.1007/s00382-017-3962-9, cited By 3.

Herrera, S., S. Kotlarski, P. M. M. Soares, R. M. Cardoso, A. Jaczewski, J. M. Gutiérrez, and D. Maraun (2018), Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution, *International Journal of Climatology*, doi:10.1002/joc.5878.

Hill, H. S., and Mjelde (2002), Challenges and opportunities provided by seasonal climate forecasts: A literature review, *Journal of Agricultural and Applied Economics*, *34*(3), 603–632, doi:10.1017/S1074070800009330.

Kharin, V. V., G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W.-S. Lee (2012), Statistical adjustment of decadal predictions in a changing climate, *Geophysical Research Letters*, *39*(19), doi:10.1029/2012GL052647.

Kintanar, R. L. (1984), Climate of the Philippines, *Tech. rep.*, PAGASA.

Koide, N., A. W. Robertson, A. V. M. Inés, J. H. Qian, D. G. DeWitt, and A. Lucero (2012), Prediction of rice production in the Philippines using seasonal climate forecasts, *Journal of Applied Meteorology and Climatology*, *52*(3), 552–569, doi:10.1175/JAMC-D-11-0254.1.

Kotlarski, S., P. Szabó, S. Herrera, O. Räty, K. Keuler, P. M. Soares, R. M. Cardoso, T. Bosshard, C. Pagé, F. Boberg, et al. (2017), Observational uncertainty and regional

climate model evaluation: A pan-European perspective, *International Journal of Climatology*, doi:10.1002/joc.5249.

Lachenbruch, P. A., and M. R. Mickey (1968), Estimation of error rates in discriminant analysis, *Technometrics*, *10*(1), 1–11, doi:10.2307/1266219.

Magnusson, L., M. Alonso-Balmaseda, and F. Molteni (2012), On the dependence of ENSO simulation on the coupled model mean state, *Climate Dynamics*, *41*(5-6), 1509–1525, doi:10.1007/s00382-012-1574-y.

Magnusson, L., M. Alonso-Balmaseda, S. Corti, F. Molteni, and T. Stockdale (2013), Evaluation of forecast strategies for seasonal and decadal forecasts in presence of systematic model errors, *Climate Dynamics*, *41*(9-10), 2393–2409, doi:10.1007/s00382-012-1599-2.

Manzanas, R., and J. Gutiérrez (2018), Process-conditioned bias correction for seasonal forecasting: a case-study with ENSO in Peru, *Climate Dynamics*, pp. 1–11, doi:10.1007/s00382-018-4226-z.

Manzanas, R., M. D. Frías, A. S. Cofiño, and J. M. Gutiérrez (2014), Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill, *Journal of Geophysical Research: Atmospheres*, *119*(4), 1708–1719, doi:10.1002/2013JD020680.

Manzanas, R., S. Brands, D. San-Martín, A. Lucero, C. Limbo, and J. M. Gutiérrez (2015), Statistical downscaling in the tropics can be sensitive to reanalysis choice: A case study for precipitation in the Philippines, *Journal of Climate*, *28*(10), 4171–4184, doi:10.1175/JCLI-D-14-00331.1.

Manzanas, R., A. Lucero, A. Weisheimer, and J. M. Gutiérrez (2018), Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts?, *Climate Dynamics*, *50*(3-4), 1161–1176, doi:10.1007/s00382-017-3668-z.

Manzanas, R., J. M. Gutiérrez, J. Bhend, S. Hemri, F. J. Doblas-Reyes, V. Torralba, E. Penabad, and A. Brookshaw (2019), Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset, *Climate Dynamics*, *53*(3–4), 1287–1305.

Maraun, D., R. Huth, J. M. Gutiérrez, D. San-Martín, M. Dubrovsky, A. Fischer, E. Hertig, P. M. M. Soares, J. Bartholy, R. Pongrácz, M. Widmann, M. J. Casado, P. Ramos, and J. Bedia (2017), The VALUE perfect predictor experiment: Evaluation of temporal variability, *International Journal of Climatology*, doi:10.1002/joc.5222.

Mason, S. J., and N. E. Graham (2002), Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, *Quarterly Journal of the Royal Meteorological Society*, *128*, 2145–2166, doi:10.1256/003590002320603584.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer, and F. Vitart (2011), *The new ECMWF seasonal forecast system (System 4)*, European Centre for Medium-Range Weather Forecasts.

Sánchez-Gomez, E., C. Cassou, Y. Ruprich-Robert, E. Fernández, and L. Terray (2016), Drift dynamics in a coupled model initialized for decadal forecasts, *Climate Dynamics*, *46*(5), 1819–1840, doi:10.1007/s00382-015-2678-y.

Shonk, J. K. P., E. Guilyardi, T. Toniazzo, S. J. Woolnough, and T. Stockdale (2018), Identifying causes of Western Pacific ITCZ drift in ECMWF System 4 hindcasts, *Climate Dynamics*, *50*(3), 939–954, doi:10.1007/s00382-017-3650-9.

Smith, D. M., R. Eade, and H. Pohlmann (2013), A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction, *Climate Dynamics*, *41*(11-12), 3325–3338, doi:10.1007/s00382-013-1683-2.

Snedecor, G. W., and W. G. Cochran (1989), Statistical methods (eight edition), *Iowa state University press, Ames, Iowa*.

Stockdale, T. N. (1997), Coupled ocean-atmosphere forecasts in the presence of climate drift, *Monthly Weather Review*, *125*(5), doi:10.1175/1520-0493(1997)125<0809:COAFIT>2.0.CO;2.

Themeßl, M. J., A. Gobiet, and G. Heinrich (2012), Empirical-statistical downscaling and error correction of regional climate models and its impact on the climate change signal, *Climatic Change*, *112*(2), 449–468, doi:10.1007/s10584-011-0224-4.

Troccoli, A., M. Harrison, D. L. Anderson, and S. J. Mason (2008), *Seasonal climate: forecasting and managing risk*, vol. 82, Springer Science & Business Media.

Vannitsem, S., D. Wilks, and J. Messner (2018), *Statistical post-processing of ensemble forecasts*, Elsevier.

Zhang, S. (2011), A study of impacts of coupled model initial shocks and state-parameter optimization on climate predictions using a simple pycnocline prediction model, *Journal of Climate*, *24*(23), 6210–6226, doi:10.1175/JCLI-D-10-05003.1.

Zhang, S., Z. Liu, A. Rosati, and T. Delworth (2012), A study of enhancive parameter correction with coupled data assimilation for climate estimation and prediction using a simple coupled model, *Tellus A: Dynamic Meteorology and Oceanography*, *64*(1), 10,963, doi:10.3402/tellusa.v64i0.10963.

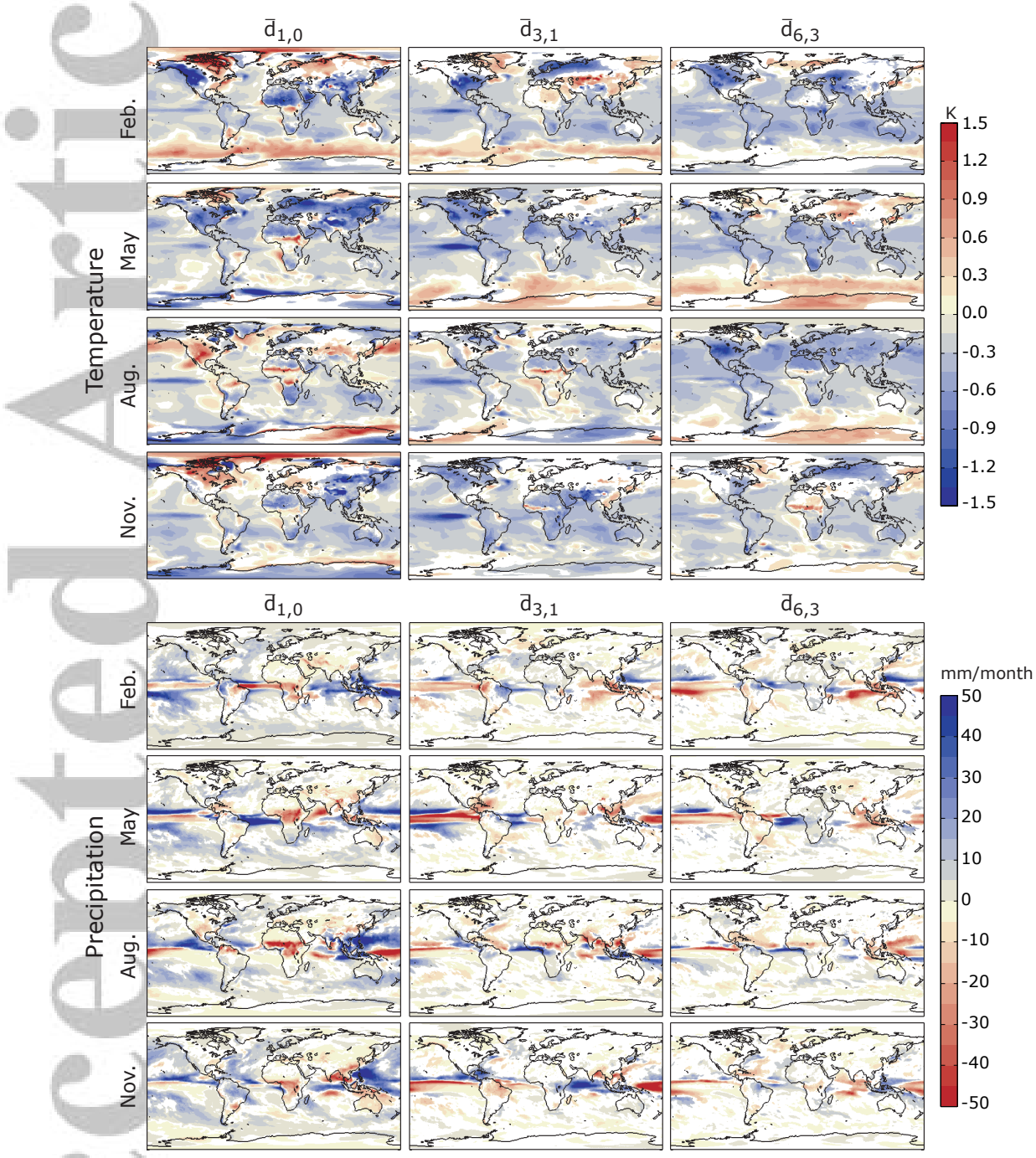
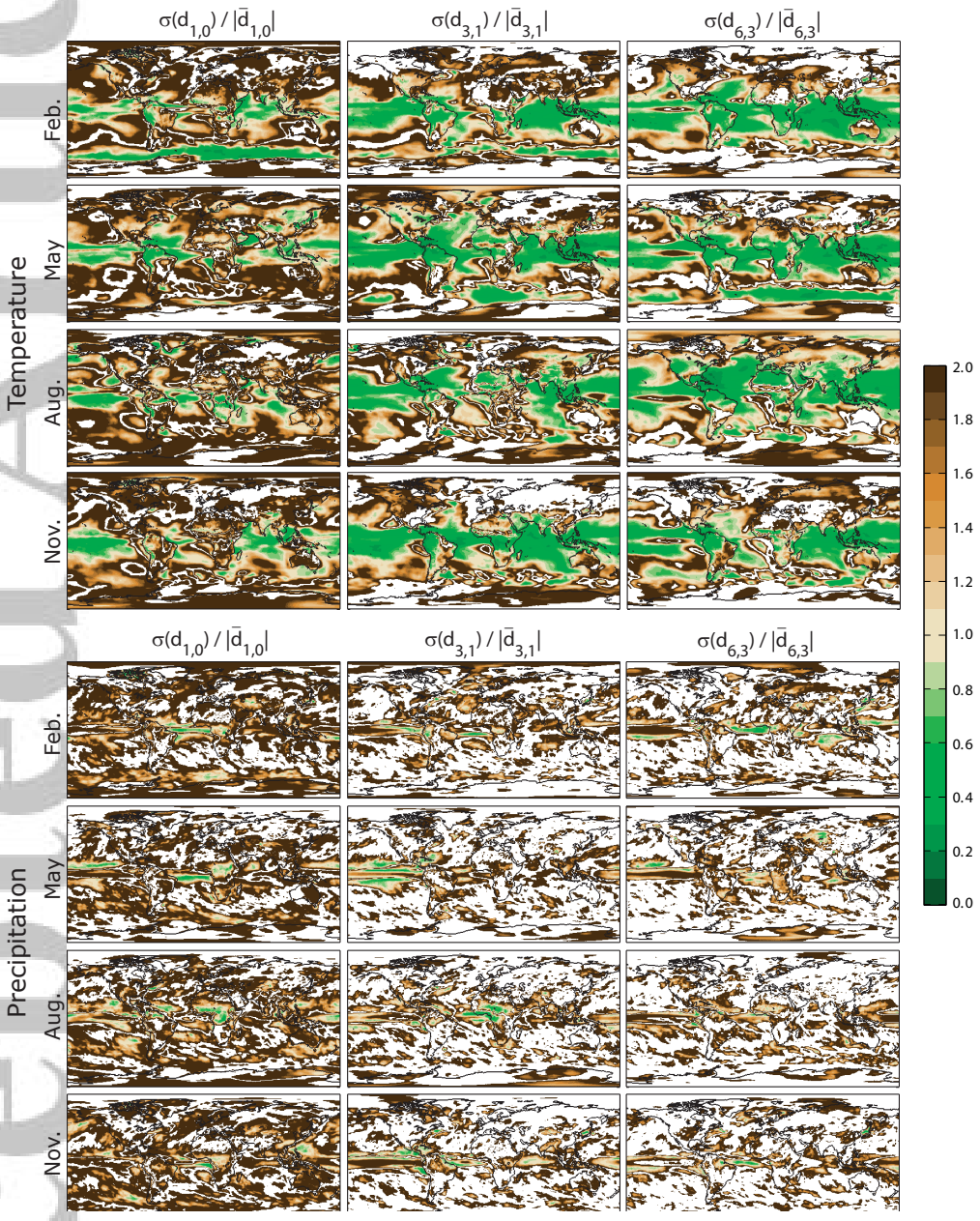


Figure 2. Drift patterns for temperature (top) and precipitation (bottom). Only values significantly ($\alpha = 0.05$) different from zero are displayed. See the text for details.



178 **Figure 3.** Inter-annual variability of the year-to-year drifts, divided by their mean absolute value —see
179 Figure 2,— for temperature (top) and precipitation (bottom). Results are only shown for regions exhibiting
180 significant drifts (colored gridboxes in Figure 2).

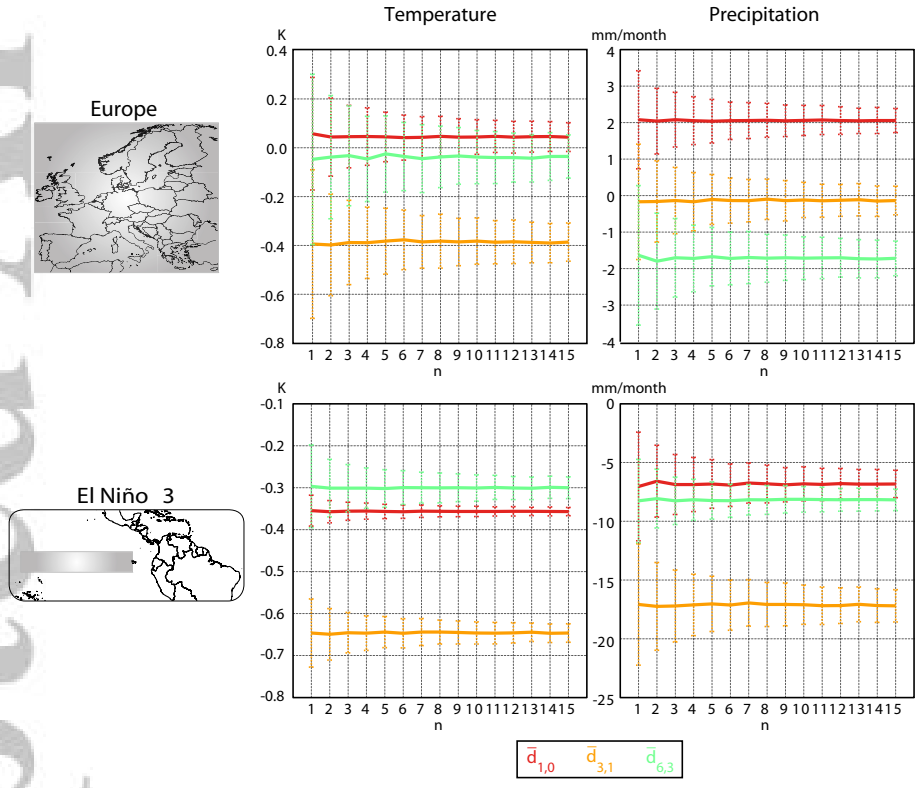


Figure 4. Model drifts, for February, as a function of the ensemble size for temperature and precipitation (left and right column, respectively), over Europe and El Niño 3 (shadowed areas in the top and bottom row, respectively). Colors correspond to different combinations of leadtimes (see the legend).

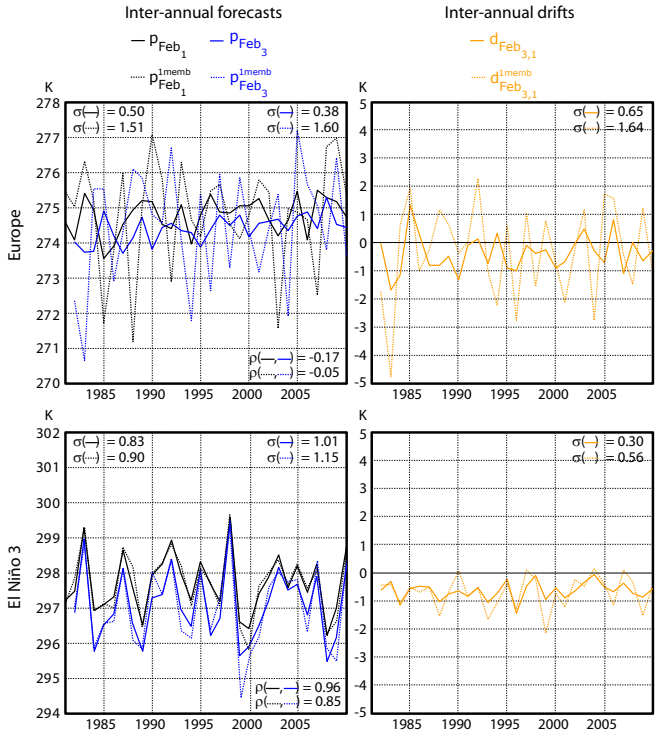


Figure 5. Left column: De-trended (Mann-Kendall test with $\alpha = 0.05$) inter-annual model predictions for temperature as given by the ensemble mean (solid lines) and a single, randomly selected, member (dashed lines), for February, over Europe and El Niño 3 (top and bottom row, respectively). Black (blue) represents one-(tree-) month lead predictions. Right column: Corresponding year-to-year drifts.

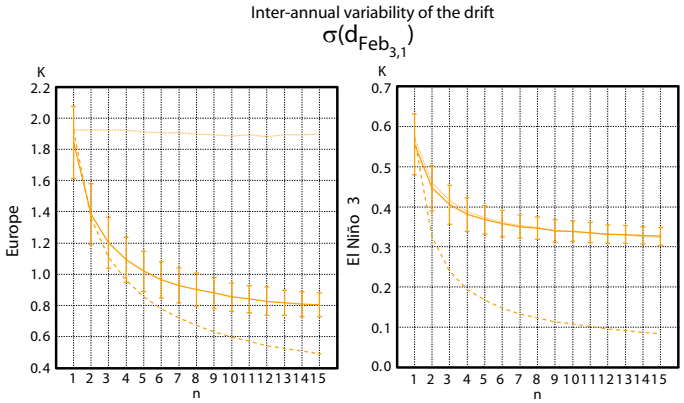


Figure 6. Inter-annual variability of the drift —in particular $d_{Feb_{3,1}}$ for temperature, as a function of the ensemble size over Europe (left) and El Niño 3 (right). Solid lines (errorbars) represent the mean value (standard deviation) obtained from 1000 bootstrapped samples. Dashed (dotted) lines draw the theoretical Equation 7 (8) for regions of low (high) model skill.

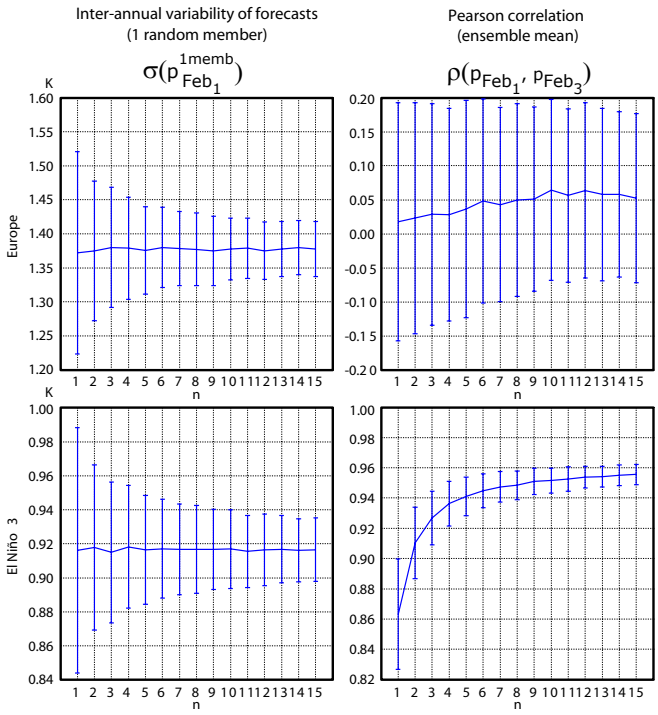


Figure 7. Dependence of the terms (1) and (2) in Equation 6 with the ensemble size (left and right column, respectively), over Europe and El Niño 3 (top and bottom row, respectively). Solid lines (errorbars) represent the mean value (standard deviation) obtained from 1000 bootstrapped samples.

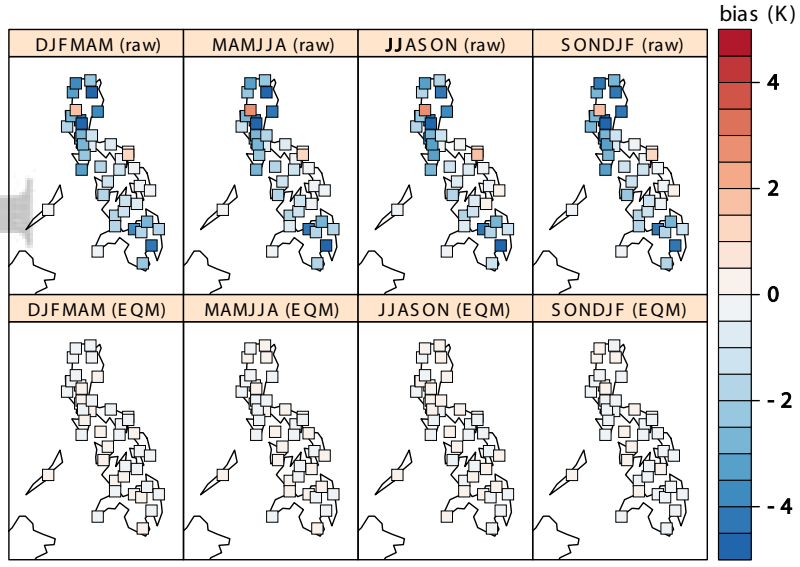


Figure 8. Mean bias obtained for temperature for the different extended seasons (in columns) at the 42 PAGASA stations, as given by the raw model forecasts (top row) and a standard implementation of the EQM method in which moving windows are not considered (bottom row).

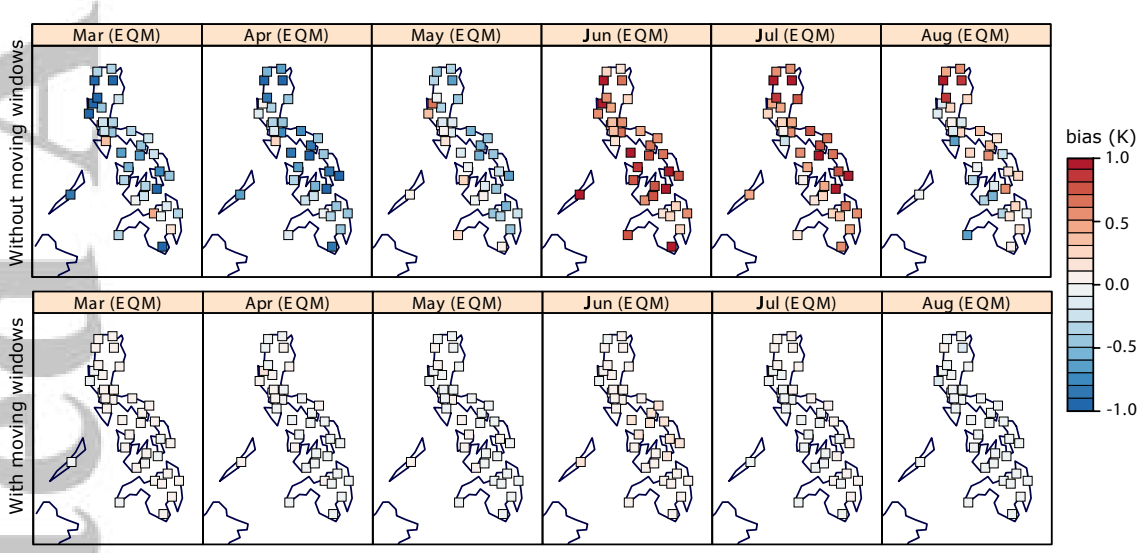


Figure 9. Mean bias obtained for temperature for each of the individual months conforming the MAMJJA season (in columns) at the 42 PAGASA stations, as given by the EQM method when moving windows are not/are considered (top/bottom row).

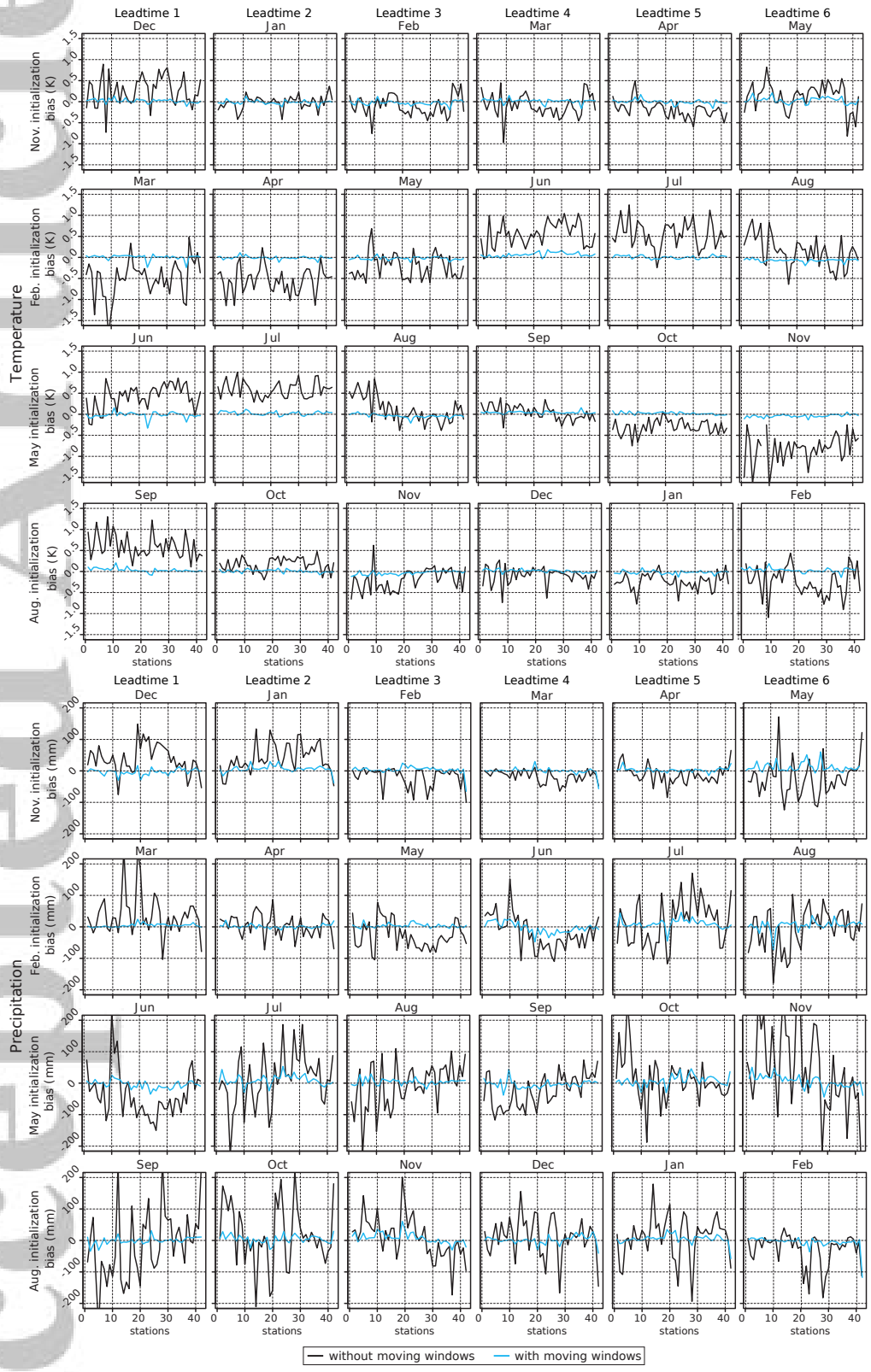


Figure 10. Mean bias obtained for temperature (top) and precipitation (bottom) for each of the individual months conforming the different extended seasons (in columns) along the 42 PAGASA stations. Black (blue) correspond to the EQM method in which moving windows are not (are) considered.

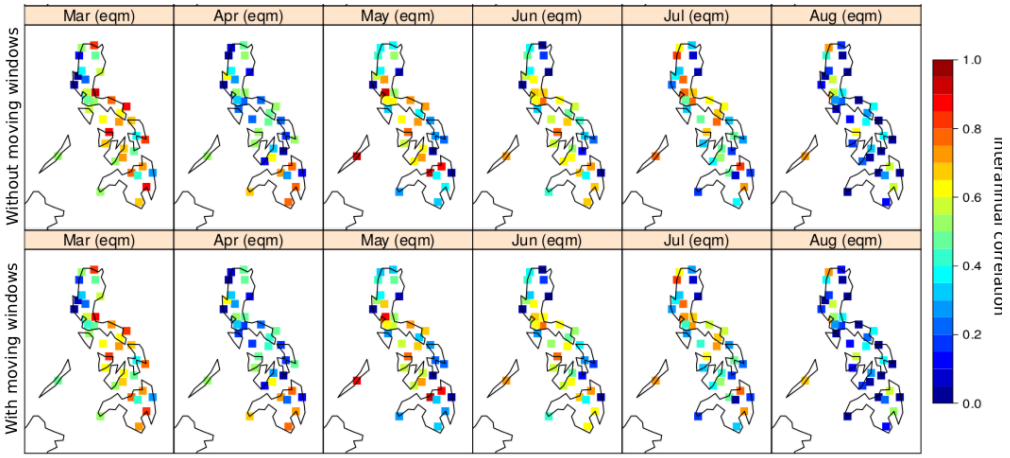


Figure 11. As Figure 9 but for interannual correlation instead of the mean bias.